On Compressed Sensing and Its Application to Speech and Audio Signals

Christensen, Mads Græsbøll; Østergaard, Jan; Jensen, Søren Holdt

Published in:
Proc. of Asilomar Conference on Signals, Systems, and Computers

Publication date:
2009

Document Version
Publisher's PDF, also known as Version of record

ON COMPRESSED SENSING AND ITS APPLICATION TO SPEECH AND AUDIO SIGNALS

Mads Græsbøll Christensen[†]*, Jan Østergaard*[‡]*, and Søren Holdt Jensen*[‡]

[†] Dept. of Media Technology
Aalborg University, Denmark
mgc@imi.aau.dk

[‡] Dept. of Electronic Systems
Aalborg University, Denmark
{jo,shj}@es.aau.dk

ABSTRACT

In this paper, we consider the application of compressed sensing (aka compressive sampling) to speech and audio signals. We discuss the design considerations and issues that must be addressed in doing so, and we apply compressed sensing as a pre-processor to sparse decompositions of real speech and audio signals using dictionaries composed of windowed complex sinusoids. Our results demonstrate that the principles of compressed sensing can be applied to sparse decompositions of speech and audio signals and that it offers a significant reduction of the computational complexity, but also that such signals may pose a challenge due to their non-stationary and complex nature with varying levels of sparsity.

1. INTRODUCTION

Sparse decompositions of signals using over-complete (or redundant) dictionaries have been used for processing of speech and audio signals for many years. Such sparse decompositions are also closely related to parametric modeling of signals, as such parametric modeling can be cast in the framework of sparse decompositions (or vice versa). Parametric models of speech and audio signals have proven to have a wide range of applications, including compression, enhancement, analysis, etc. In fact, audio coders based on these principles have already been a part of the MPEG audio coding standards for a long time. Curiously, the fields of parametric modeling, including estimation theory, and sparse decompositions have largely developed independently, ignorant of each others' important ideas and results.

Recently, important new theoretical advances in what has been dubbed compressive sampling or compressed sensing (CS) have been made [1, 2, 3] and has spawned a flurry of activity in research on this topic. The basic results are, basically, that under some fairly weak conditions, signals that are composed as linear combinations of few linearly independent vectors need only to be sampled at a low rate to facilitate a high quality reconstruction. Here, few means that the number of basis vectors is small relative to the number of samples. More specifically, if a signal is composed of a linear combination of T vectors, we can reconstruct the signal using κT samples (where κ is a small positive integer) formed as random linear combinations of the, say, N original samples . It is then clear that if κT is much smaller than N, we have achieved a compression of sorts, a compression that can be implemented directly in the sampling, provided that the level of sparsity is known a priori. Hence the name compressive sampling.

Until now, this principle has only been applied to speech and audio signals in the context of linear predictive coding of speech, where the sparsity is in the residual domain [4, 5]. In this paper, we consider the application of the principles of compressed sensing to speech and audio signals in the context of sparse decompositions based on redundant dictionaries. We consider the possible advantages of doing so and the related design issues that must be addressed and possible caveats. Moreover, by our choice of dictionary, namely windowed complex exponentials, we demonstrate that the principles also apply to parametric modeling of speech and audio signals.

The remaining part of this paper is organized as follows. In the next section, Section 2, we cast the problem of sparse decompositions within the compressed sensing framework and discuss the basic results of compressed sensing. Then, in Section 3, we discuss its application to speech and audio signals and the related design issues. Finally, we provide some examples of the application of the sparse decompositions and compressed sensing to real speech and audio signals in Section 4 before concluding on our work in Section 5.

2. SPARSE DECOMPOSITIONS AND COMPRESSED SENSING

We will start out by first introducing the sparse decomposition before considering its combination with CS. The sparse decomposition problem can be defined as follows. Given a segment of a signal $\mathbf{x} \in \mathbb{R}^N$ and a fat matrix $\mathbf{Z} \in \mathbb{C}^{N \times F}$ containing the dictionary and we seek to find a sparse coefficient vector $\mathbf{c} \in \mathbb{C}^F$ with $F \gg N$ that recovers $\mathbf{x}$ exactly,

i.e.,

$$\mathbf{x} = \mathbf{Z}\mathbf{c}, \tag{1}$$

or approximately. To do this, we need to introduce a sparsity metric on $\mathbf{c}$. A commonly used measure for this is the vector 1-norm, denoted $\|\cdot\|_1$, which can be related to the number of non-zero coefficients under certain technical conditions. The vector $\mathbf{c}$ is said to be T-sparse if it contains T non-zero coefficients and we will here use this synonymously with the term sparsity. We can now pose the sparse decomposition problem as the following:

$$\begin{aligned} &\text{minimize } \|\mathbf{c}\|_1 \\ &\text{s.t.} \quad \mathbf{x} = \mathbf{Z}\mathbf{c}. \end{aligned} \tag{2}$$

As an example, consider the signal produced by a pitched instrument. Such a signal can be well-modeled by a finite sum of sinusoids. In this case, the dictionary consists of complex sinusoids and the entries in $\mathbf{c}$ are their complex amplitudes. We then seek to describe the signal using as few non-zero complex amplitudes as possible, corresponding to selecting a low-order sinusoidal model.

The problem in (2) is what is referred to as a second order-cone program (SOCP) in the convex optimization literature, and it can be solved efficiently using standard methods. It should be stressed that it cannot be cast as a linear program and it is therefore not *basis pursuit* as originally introduced in [6]. This is because the 1-norm of the stacked real and imaginary parts of vector $\mathbf{c}$ is not the same as the 1-norm of $\mathbf{c}$.

The main idea of CS [1, 2, 3] is to map the observed signal $\mathbf{x}$ to a lower-dimensional vector $\mathbf{y} \in \mathbb{R}^K$ with $K < N$ via a fat so-called measurement matrix $\mathbf{\Phi} \in \mathbb{R}^{K \times N}$ by the following transformation:

$$\mathbf{y} = \mathbf{\Phi}\mathbf{x}. \tag{3}$$

Similarly, the measurement matrix is also multiplied on to $\mathbf{Z}\mathbf{c}$, and we can write the sparse decomposition problem as

$$\begin{aligned} &\text{minimize } \|\mathbf{c}\|_1 \\ &\text{s.t.} \quad \mathbf{\Phi}\mathbf{x} = \mathbf{\Phi}\mathbf{Z}\mathbf{c}, \end{aligned} \tag{4}$$

or, introducing $\mathbf{Q} = \mathbf{\Phi}\mathbf{Z}$, as

$$\begin{aligned} &\text{minimize } \|\mathbf{c}\|_1 \\ &\text{s.t.} \quad \mathbf{y} = \mathbf{Q}\mathbf{c}. \end{aligned} \tag{5}$$

The key point of CS theory is that under certain conditions, namely an appropriate choice of measurement matrix $\mathbf{\Phi}$ [2, 3, 7], solving (2) will result in a solution vector $\hat{\mathbf{c}}$ identical to that of (5). Therefore, reconstruction of the signal using $\hat{\mathbf{c}}$ obtained from (5) will reconstruct not only $\mathbf{y}$ as in constraints of (5), but also $\mathbf{x}$ exactly as $\mathbf{x} = \mathbf{Z}\hat{\mathbf{c}}$ when $\mathbf{c}$ is sparse. We note that while many of the proofs and conditions have been stated in the literature for orthogonal bases, the principles apply also for frames [1], as is the case considered here.

In the CS literature, much emphasis has been put on the ramifications of these results for simplifying sensors and A/D converters. From the above discussion, it is, however, clear that also in traditional applications of sparse decompositions, i.e., to signals that have already been sampled, there is a possibly huge benefit of using the principles of CS, namely that the problem in (5) is of a possibly much lower dimensionality than the original sparse decomposition problem in (2). More specifically, the problem in (2) involves solving for F variables subject to N constraints while (5) involves also F variables, but only K constraints. If K is much smaller than N, the savings in computational complexity is potentially huge as problems of the forms considered here generally have cubic complexity. CS may therefore be used as a pre-processor for the many applications of sparse decompositions we have seen through the past decade. That the complexity involved with solving problems of the form (2) has been a concern and a limiting factor can be witnessed by the wide spread use of approximate solutions obtained using greedy methods like matching pursuit.

3. DISCUSSION

Next, we will discuss some important issues in applying CS and sparse decompositions to speech and audio signals.

A) Dictionary Complex sinusoids have been reported to work well for sparse decompositions for a wide range of applications. They are, however, also well-known to perform poorly for modeling transient phenomena, even if these are modulated sinusoids, and also stochastic signal components. The solution to these problems are generally composite dictionaries (or unions of bases), where also modulated sinusoids and even time-domain Kronecker delta functions are included. Also, voiced speech signals can be modeled not only as sparse in the frequency domain, but also as sparse in the residual domain after linear prediction has been applied [4, 5]. This complicates matters somewhat for CS as the measurement matrix should be chosen such that it is incoherent with the dictionary. Fortunately, random measurement matrices can generally be expected to have a low coherence with both time-domain spikes and complex sinusoids. Furthermore, random measurement matrices that are universally applicable regardless of the type of dictionary can be constructed [7].

B) Sparsity Another issue with speech and audio signals is that the sparsity of such signals may vary greatly over time; a low piano note may contain many partials while a glockenspiel may contain only a few. At one particular time instance of a piece of music, a single instrument playing just a single note may be present while at other times, mul-

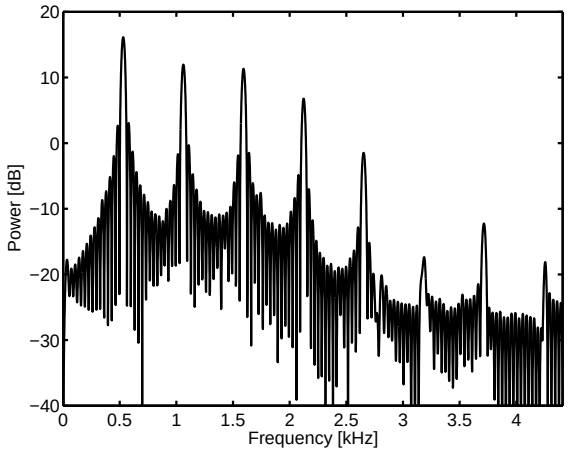

Fig. 1. Power spectrum of trumpet tone.

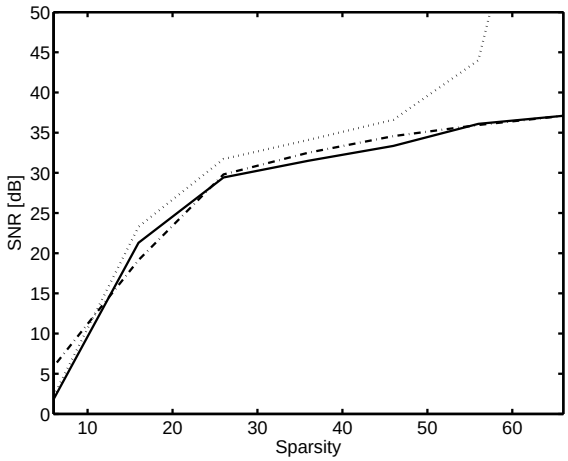

Fig. 2. Reconstruction SNR in dB as a function of the assumed sparsity T (dotted) with only the assumed number of non-zero coefficients retained with CS (solid) and without (dash-dotted).

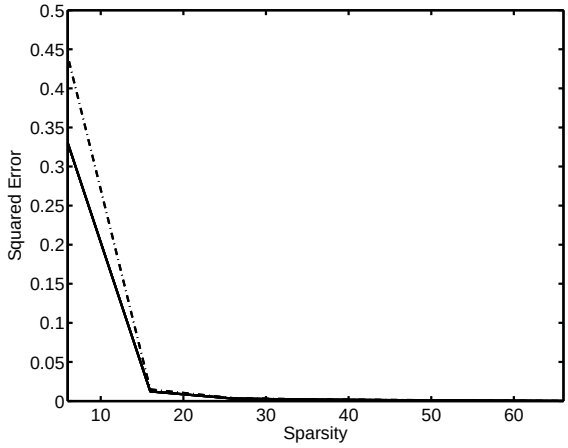

Fig. 3. 2-norm of difference between coefficients found with and without CS (dash-dotted) and when only retaining the assumed number of non-zero coefficients (solid).

tiple instruments playing multiple notes may be playing at the same time. This means that we would have to either a) bound the sparsity of the signal by worst-case considerations, or b) somehow estimate the sparsity of the signal over time. The feasibility of the former approach can be illustrated by considering voiced speech. The fundamental frequency of speech signals can be be between 60 and 400 Hz. This means that there can be at most 66 harmonics for a signal sampled at 8000 kHz. The latter approach would have to be computationally simple or it would otherwise defeat the purpose of the CS. There is another aspect that should be taken into account, though. It is generally so that the larger the dictionary, the sparser a coefficient vector one can obtain. This is easy to explain. Suppose a signal contains a single sinusoid having a frequency in-between the frequencies of two vectors in the dictionary, then the contribution of that single sinusoid will spread to several coefficients. The expected sparsity of the coefficient vector $\mathbf{c}$ and thus the number of samples K required for reconstruction is therefore not only a function of the signal, but also the dictionary. It is thus more likely that a single dictionary element will match the signal (or part thereof) if the dictionary is large.

C) Noise First, let us elaborate on what we mean by noise: we mean all stochastic signals contributions, everything that cannot easily be modeled using a deterministic function. Stochastic signal components are inherent and perceptually important parts of both speech and audio signals. Stochastic components occur in speech signals during periods of unvoiced speech or mixed excitation where both periodic and noise-like contributions are present. Similar observations can be made for audio signals. This means that, depending on the application, CS may not be entirely appropriate. On the other hand, if only the tonal parts of the signal are of interest, then it may yet be useful. It should also be noted that the characteristics of the noise in speech and audio signals are time-varying, and it can often be observed to vary faster than tonal parts of audio signals. This again implies that it may be difficult to determine the required number of samples a priori and the expected reconstruction quality in LASSO-like reconstructions [8].

4. SOME EXAMPLES

We will now present some results regarding the application of CS to speech and audio signals. First, we will conduct an experiment on a single segment of data, namely 60 ms of a stationary trumpet tone whose power spectrum is shown in Figure 1. For visual clarity, this signal has been down-

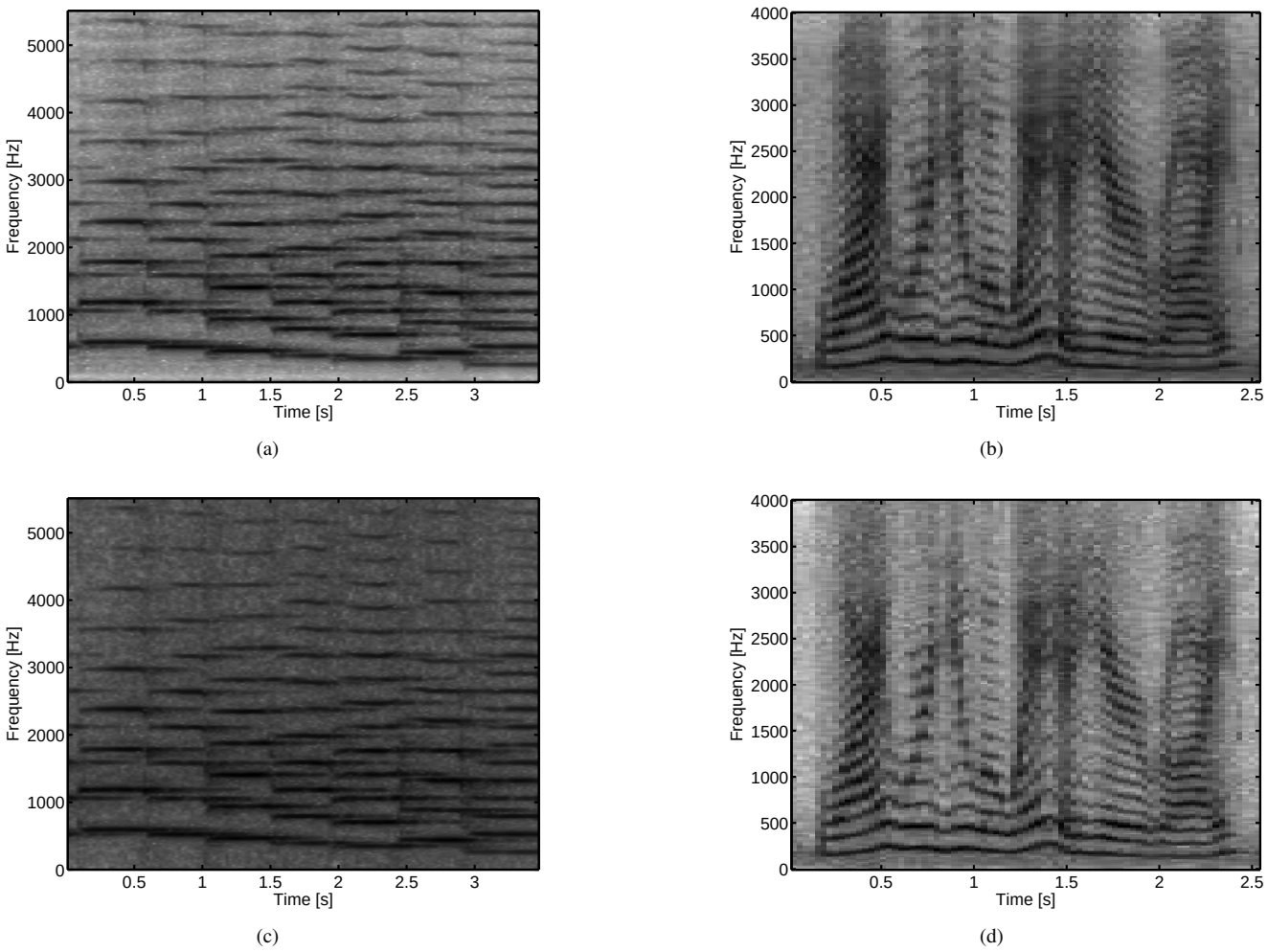

Fig. 4. Spectrograms of original signals, trumpet (a) and voiced speech(b), and the same signals reconstructed using CS with 60 non-zero coefficients for each 40 ms segment (c) and (d).

sampled from 44.1 kHz by a factor 4. As can be seen, the spectrum is highly sparse with only 8 harmonics being present (note that because the signal is real, 16 non-zero coefficients are needed at a minimum to capture these). We have performed sparse decomposition of this signal with and without CS using a dictionary composed of $F = 4096$ complex sinusoids having frequencies uniformly distributed between 0 and 2π. If such a signal lends itself to CS, we should be able reconstruct the signal at the same SNR in both cases. When using CS, one must choose how many samples to retain. As rule of thumb four times as many samples as the number of non-zero coefficients should be used [8], i.e., $K = 4T$, and this is what we do here (for more on this, see [9]. Furthermore, a measurement matrix constructed from independent and identically distributed zero-mean Gaussian variables is used throughout these experiments.

In Figure 2, the reconstruction SNR (in dB) is shown as a function of the assumed sparsity, i.e., the number of non-zero coefficients T in $\mathbf{c}$. First, the SNR for the solution obtained using (5) (dotted) is depicted. For comparison, (2) results in a reconstruction SNR of more than 200 dB. Also shown is the reconstruction SNR when only the T largest (in magnitude) coefficients are retained for the two solutions obtained using (5) (solid) and (2) (dash-dotted). It can be seen that when sparsity is enforced, the two method perform similarly, only (5) is solved much faster as the problem is much smaller. This suggests that CS indeed retains the information required to reconstruct the sparse part of the signal, i.e., the sinusoids, and this appears to be the case regardless of the assumed level of sparsity in the sense that even if the assumed level of sparsity is below the actual level, the sparse decomposition with CS still works approximately as well as it would have without it. This is an

important observation as the level of sparsity cannot generally be known a priori for speech and audio signals. It means that we basically do no worse than we were doing with the sparse decomposition in the first place. To investigate whether the same coefficient vector is obtained using the two approaches, the 2-norm of the difference between the obtained vectors are shown in Figure 3 for the full vectors obtained from (5) and (2) (dash-dotted) and when only the assumed number of non-zero coefficients T is retained (solid). Again, this confirms that CS is applicable to the signal in questions and that the difference between the two solutions tends to zero for a sufficiently high T and K and that the difference is smaller for the largest coefficients.

Next, we will process and reconstruct a longer fragment of a trumpet signal sampled at 44.1 kHz whose spectrogram is shown in Figure 4(a) (for visual clarity, only the lower part of the spectrum is shown, although the signal does have harmonics extending beyond the visual part). We do this as follows. We process the signal in 40 ms segments with 50 % overlap using a dictionary composed of $F = 4096$ windowed complex sinusoids, i.e., Gabor-like frames, and construct a measurement matrix as in the previous experiment. The signal is synthesized using overlap-add. We here assume a sparsity of $T = 60$ non-zero coefficients and use only the T largest coefficients in reconstructing the signal and use 4 times as many samples in the CS process, i.e., $K = 240$ samples. This means that while the original sparse decomposition problem in (2) is solved subject to 1764 constraints, we have reduced this to just 240 using CS, i.e., by a factor more than 7. Considering that solving such problems involves algorithms of cubic complexity, we can therefore expect a significant reduction in computation time and this has been confirmed by our simulations. In Figure 4(c), the reconstructed signal is shown, and it can be seen that the harmonic parts of the spectrum have indeed been recovered.

A similar experiment was carried out for a voiced speech signal sampled at 8 kHz. The spectrogram of the original signal is shown in Figure 4(b) while the reconstructed signal is shown in Figure 4(d). The signal was processed as before, with 40 ms segments and 60 non-zero coefficients assumed in the CS and the reconstruction. As can be seen, the part of the speech signal that is sinusoidal has been captured, but it can also be observed that some of the upper parts of the spectrum, that have a more stochastic nature, have been lost. Our results clearly confirm that the principles of CS can be used for periodic signals such as voiced speech and tonal audio signals.

5. CONCLUSION

In this paper, we have considered the application of the principles of compressed sensing to speech and audio signals. More specifically, we have done this in the context of sparse decompositions based on dictionaries comprised of windowed complex exponentials. We have argued that compressed sensing may serve as a pre-processor for sparse decompositions as the complexity of solving the involved convex optimization problems is greatly reduced in the process. Furthermore, our results demonstrate that sparse decompositions work equally well with and without compressed sensing regardless of the assumed level of sparsity. This is an important observation as the level of sparsity cannot be known a priori and may vary over time for speech and audio signals. This basically means that sparse decompositions with compressed sensing works no worse than sparse decompositions did in the first place.

6. REFERENCES

[1] D. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52(4), pp. 1289–1306, apr 2006.

[2] E. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52(2), pp. 489–509, feb 2006.

[3] E. Candes and T. Tao, "Near optimal signal recovery from random projections: Universal encoding stragies?," *IEEE Trans. Inf. Theory*, vol. 52(12), pp. 5406–5425, dec 2006.

[4] D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen, and M. Moonen, "Sparse linear predictors for speech processing," in *Proc. Interspeech*, 2008.

[5] T. V. Sreenivas and W. B. Kleijn, "Compressive sensing for sparsely excited speech signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2009, pp. 4125–4128.

[6] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, pp. 33–61, 1996.

[7] R. Baraniuk, M. Davenport, R. Devore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constr. Approx*, vol. 2008, 2007.

[8] E. Candes and M. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25(2), pp. 21–30, Mar. 2008.

[9] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso)," *IEEE Trans. Inf. Theory*, vol. 55(5), pp. 2183–2202, May 2009.